DOCUMENT RESUME

ED 414 308 TM 027 775

AUTHOR	Hill, Richard
TITLE	States Set Common Standards, IF...
PUB DATE	1997-06-18
NOTE	8p.; Paper presented at the Annual Assessment Conference of the Council of Chief State School Officers (Colorado Springs, CO, June 1997). Printed on yellow paper.
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Academic Achievement; Comparative Analysis; *Elementary Secondary Education; Geographic Regions; National Surveys; *Regional Characteristics; *Standards; *State Programs; Testing
IDENTIFIERS	*National Assessment of Educational Progress; Standard Setting; *Trial State Assessment (NAEP)

ABSTRACT

Whether differences in the standards states have set can be explained by something other than regional differences is explored. In addition, a way in which standards can be compared is defined, and the standard of proficiency that seems to be widely shared across the country is illustrated. The Trial State Assessment (TSA) data from the National Assessment of Educational Progress (NAEP) are an excellent tool for answering the question of differences in standards. In looking at the TSA data, it is necessary to consider the following aspects of testing: (1) testing equivalent populations; (2) testing under equivalent conditions; (3) the year of the testing program; (4) having similar frameworks and the opportunity to learn; (5) testing at different grade levels; (6) testing in different content areas; and (7) using different years of NAEP data. A look at all these factors makes comparisons complicated, yet some clear and strong patterns emerge. The states tend to cluster in three groups. The first, with standards generally far below those of the NAEP, tend to be in the South. The second, with standards near, but still below the NAEP, is mostly made up of Northeast and North Central states, while the third, with standards somewhat above the NAEP, are states for which Advanced Systems in Measurement and Evaluation, Inc. is the standard-setting contractor. Clear regional differences in standards are noted. (SLD)

States Set Common Standards, IF ...

Richard Hill
Advanced Systems

Paper Presented at the 1997 CCSSO Annual Assessment Conference
Colorado Springs
June 18, 1997

Introduction

An impetus for this paper was a discussion I had with Mark Musick of the Southern Regional Education Board, in which he told me about his findings that standards varied greatly from state to state. In contrast, I had found substantial commonality of standards in the states for which Advanced Systems was the contractor. Musick originally had published his findings in a paper entitled, "Setting Education Standards High Enough." An update of that paper was summarized in *Education Week*. The update provided a list of over a dozen states in alphabetical order, and showed the percentage of students who were meeting the state's standard for proficiency in grade 3 or 4 in mathematics, and grade 7 or 8 in reading in 1994-95, and compared that to the percentage of students in the state who had scored at the Proficient level on NAEP. At the extremes in that chart were Louisiana, where 88 percent of the students passed the state's criterion for proficient, in contrast to 15 percent passing NAEP's criterion; and Delaware, where only 11 percent passed the state's criterion for proficiency, but 23 percent passed NAEP's. Clearly, the level of achievement need to be labeled as "Proficient" (or some term equivalent to that) within these two states were quite different.

In contrast, Advanced Systems had been involved in setting standards in several states for which it was the contractor. As my co-presenters in this session, Gayle Potter from Arkansas and Ed Reidy from Kentucky, will show, those standards were readily replicated, even when the groups setting the standards were quite small. Thus, our experience has been that standards are quite stable, and that even very small and very different groups tend to come to similar conclusions about what is proficient and what is not.

When I pointed out this commonality to some people who have spent some time thinking about issues in standard setting, I generally got the answer that the one element in common behind the similarities in Kentucky, Maine, New Hampshire and Arkansas was Advanced Systems—that we were doing something unique that was forcing the standards to be in common. That is, those states might not have arrived at the same standards if someone else had led the standard setting process. That answer never was satisfying, because it presumed all staff from Advanced Systems had used the same process. In fact, those standards were set over a period of several years (from 1992 to 1996) by different staff members, and our approach changed as we learned from earlier experiences and different people applied those lessons in different ways. While there clearly were common elements in the procedures used in those states, such as the heavy reliance on extended open-response questions, there were many differences as well. For example, Arkansas had completed two very different studies. One was loosely controlled but included almost a thousand people, while the other was tightly controlled but was limited to a small group of hand-selected people. The standards set by these two very different processes resulted in highly similar standards for two of three content areas, and a clear understanding of why the results in the third area were different.

Therefore, the primary purpose of this paper is to determine whether differences in the standards states have set can be explained by something other than regional differences. Secondary purposes include defining a way by which standards can be compared, showing the robustness of that method, and illustrating the standard of proficiency that seems to widely shared across the country.

How Can You Tell If States Are Setting Comparable Standards?

As Musick pointed out, the Trial State Assessment data from the National Assessment of Educational Progress are an excellent tool for answering this question. There are some downsides to the TSA data: it's only available for a limited number of grade levels and content areas, not all states participated, and one has to believe that NAEP measures basically the same content standards as the statewide tests for the participating states. Despite these limitations, there is a great deal of value to the NAEP TSA data.

Musick did not do any manipulation to the NAEP data. He straightforwardly compared the percentage of students passing the state's standard to the percentage passing NAEP's standard, in part because his purpose was to simply show that there were great disparities among states' standards. However, one purpose of this paper is to quantify these disparities, so an approach that can put them all on a common scale iss needed.

This approach depends on another set of information provided by NAEP—the NAEP scaled scores representing the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles in each state. If one converts those percentiles to z-scores, the data become quite linear and it is possible to make accurate interpolations. Therefore, the process used in this paper was, in essence, equipercentile equating. That is, if we knew that for any given state, x percent of the students passed the state's test in its first year of administration, then we interpolated the NAEP data to determine the NAEP scaled score that was surpassed by x percent of the students on that test.

For example, according to Musick's paper, 39 percent of eighth grade students in New Jersey showed "clear competence" in mathematics in 1994-95; in other words, the passing score on New Jersey's test that year was the 61st percentile. In 1992, on the TSA, the 50th percentile for New Jersey was a scaled score of 273; the 75th percentile was 297. The 50th percentile in a normal distribution is a z-score of 0.00; the 61st percentile is a z-score of .28; and the 75th percentile is a z-score of .67. Therefore, the interpolated estimate of the 61st percentile on the TSA is 273 + (.28/.67) * (297 - 273), or 283. Thus, the 61st percentile student in grade 8 in New Jersey was the lowest scoring student to pass their mathematics assessment in 1994; the 61st percentile student in grade 8 in New Jersey had a NAEP scaled score of 283 in the TSA grade 8 mathematics assessment in 1992. Subject to some limitations that will be discussed below, the best guess that one could make, from the information given, is that the NAEP scaled score most equivalent to New Jersey's grade 8 math standard is 283.

Issues Affecting Comparability

There are several reasons why the estimation made in the above section might be inaccurate. The following is a list of some of them, along with a discussion on the likely size of the effect and means to minimize them.

Testing Equivalent Populations: Whenever testing is done, some students are left out of the results. The issue might be as small, and probably as random, as excluding students who are absent from school because there is no make-up testing; or it might be potentially large, as excluding all students who are likely to perform poorly because there are significant stakes associated with the test and the school has the opportunity to exclude those students who will adversely affect their scores.

NAEP has its own exclusion rules and reasons for not testing students. For example, students with special needs are excluded if so designated by the school. Since NAEP results are not reported at the student level, one presumes that there is minimum incentive for a school to exclude a large percentage of its students. Nonetheless, the NCES standard is for a minimum participation rate of 85 percent, which means that as many as 15 percent of students could be missing from the results. For the 1992 assessment, for example, 6 percent of the 9-year olds, 9 percent of the 13-year-olds, and 17 percent of the 17-year-olds did not complete the assessment, and therefore were not included in the results. Contrast that statistic to Kentucky's, where less than 1 percent of the students were excluded from reporting (those not tested are reported as Novice in Kentucky, so its reported statewide percentages of Proficient students essentially uses the number of students enrolled as its denominator).

Obviously, the more students tested, the lower a state's results are likely to be, and thus, the poorer it is likely to score on its own test relative to NAEP. If two states set exactly the same standard, but one tested more students than the other, then that state will pass fewer students and therefore its standard will *appear* to be higher. This effect will vary from state to state, but it can be directly examined for each state by obtaining the necessary information about exclusion rules and their enforcement, as well as by knowing testing rates, which are the result of those rules.

Testing under Equivalent Conditions. One presumes that, within limits, students that are more highly motivated will score higher on tests than students who are not as motivated. The motivation for students to score well on NAEP is quite limited; there are no stakes for them or their school. Contrast this to a state where passing the test is a high school graduation requirement, or where the jobs of teachers potentially are on the line. Under such circumstances, it can be presumed that there is much more pressure on students to do well. This effect is likely to have more impact on results for higher grades than for lower grades; one presumes that, without stakes, fourth graders give a better effort on a test than do twelfth graders.

If higher motivation leads to higher test scores, then if two states have the same standard, but one administers its tests under conditions of higher motivation, its students will score better, and therefore its standard will *appear* to be lower.

Year of the Testing Program. This is perhaps one of the more important issues to consider that is often overlooked. Kentucky's experience in this area is not uncommon. When its new testing program, KIRIS, began, a small percentage of students were scoring at the proficient level. As students and teachers became familiar with the testing program, and as curriculum designed to improve scores on the new testing program became implemented, the percentage of students designated as Proficient rose dramatically. There were increases in NAEP scores, too, to be sure, but those increases were quite modest in contrast to those on the KIRIS tests. Thus, of all the variables that are easy to control, determining the year of the testing program in which the equivalence is done is the most important. *Whenever determining the equivalence of a state's testing program scores to those of NAEP, always use the first year of the state's testing program.* That is the time when the preparation of students and the alignment of the curriculum to the state's test is going to be most like its preparation and alignment for NAEP—i.e., minimal. Given the dramatic increases in test scores that can occur in the first few years of a testing program, it is easy to see why this variable is so important. Using the first year's data will provide the most appropriate equivalences; data from subsequent years is likely to be inflated, and therefore make the state's standard *appear* to be lower.

Having similar frameworks and opportunity to learn. The more a state's framework is similar to the ones used to develop NAEP, the more students in a state have had similar opportunities to learn tasks on the two tests, and the more a state's test is like NAEP, the more likely it is that these equivalences will be valid. When states use strategies on their tests that are novel, it is likely that students will be more prepared for NAEP than they will be for their own state's test, until the curriculum and instruction in the state catch up to the new design. When states' tests make extensive use of open-response questions, for example, it is likely that students will perform more poorly on such tests until they have had an opportunity to learn how to effectively communicate their knowledge on such tests. When a state uses a novel test design, it is likely that the impact will be, at least initially, to make the state's standard *appear* to be higher.

Testing at different grade levels. While most states have aligned their grades of testing to those of NAEP, some test at other than grades 4, 8 and 12. If one knows, say, that the median student in the state scored at the 45th percentile on NAEP in grade 4, what is the likelihood that the median student in the state would have scored around the 45th percentile on NAEP if the test had been given at grade 3 or grade 5 instead?

The question can be answered for mathematics only, since that is the only content area for which TSA has been administered at more than one grade. In 1992, the correlation between state means at grades 4 and 8 was .96; in 1996, the correlation was .92[1]. So, at least between those grades, the relative rankings of states remains quite stable. Whether it is reasonable to extrapolate beyond those grades, and in reading, as I have done later in this paper, is debatable.

Testing in different content areas. "Guesstimating" what a state's score would be in other than grade 4 reading and grade 4 and 8 mathematics is going to be complete conjecture, since there is no information to provide in other content areas. However, it is worth noting that the correlation

[1] These correlations include only states, not the territories and Washington, D.C. Those additional areas have extreme scores. As a result, including them in the calculations would tend to inflate the size of these correlation coefficients and be misleading.

between states' median scores in reading and mathematics in the 1992 TSA were quite high. Again, including only the 41 states that participated in the assessment (i.e., excluding the territories and Washington, D.C), the correlation between states' median scores in grade 4 reading and grade 4 mathematics was .94. Even the correlation between the median scores in grade 4 reading and grade *8* mathematics was .90. Thus, a state's relative ranking in any grade level and content area is likely to be quite stable.

Using different years of NAEP data. The first TSA was held in 1992, and the second in 1996. If a state were to start a new testing program in 1998, for example, would it be acceptable to establish equivalence on the basis of the 1996 data?

The answer appears to be clearly in the affirmative, with one *caveat*. States tend to rank very similarly across grades, across reading and mathematics, and across time. The correlation between grade 4 math in 1992 and 1996 was .90; for grade 8, the correlation was .97. Looking at the more extreme example—grade 4 math—one sees that over the four years, the average increase in test scores was 3 points. However, two of the 37 states participating in both years increased their scores by 11 points while one dropped by three points. The standard deviation of student scores was about 32 points, so a gain of 11 points is about 1/3 of a standard deviation. Given the apparent gross discrepancies in standards that precipitated this paper, that is a small effect. However, if one were trying to estimate scores with a high degree of accuracy, the example of grade 4 mathematics shows that scores sometimes do change over time.

How Comparable Are States' Standards?

Given all the caveats above, it is clear that one needs to read the data in Table 1 with some caution. Surely almost every conceivable combination of testing rates, motivation, year of testing program and novelty of assessment design is included in these numbers. Despite all these concerns, however, some clear patterns emerge. Also, these patterns are so strong that even if all the issues discussed above vary systematically from state to state, they would not explain away all the differences.

The states tend to cluster into three groups. The first group, with standards generally far below those of NAEP, tend to be those from the South. The second group, with standards near, but still below, those of NAEP, includes mostly states from the Northeast and North Central portion of the country. The final group, all with standards somewhat above those of NAEP, are the states for which Advanced Systems is the contractor (although Advanced Systems was not involved in the standard setting done in Delaware). There are only two exceptions to this grouping: Wisconsin and Rhode Island (in mathematics only). Thus, there are clear regional differences in these results.

Table 1

NAEP Scaled Score that Is Equivalent to States' Standards

State	Grade 4 Reading		Grade 8 Math	
	Percent Passing	Equivalent NAEP Scaled Score	Percent Passing	Equivalent NAEP Scaled Score
Louisiana	88	165	80	220
Wisconsin	88	188	--	--
Georgia	67	198	83	225
North Carolina	65	200	68	241
South Carolina	82	179	68	242
Tennessee	62	204	--	--
Oklahoma	--	--	70	252
Rhode Island	65	206	62	256
Texas	79	186	56	258
Indiana	66	211	41	262
Michigan	--	--	55	263
Maryland	34	230	42	273
Connecticut	48	228	47	278
New Jersey	--	--	39	283
Idaho	--	--	35	287
NAEP		243		294
New Hampshire	26	249	--	--
Kentucky	7	261	14	298
Delaware	11	257	13	302
Maine	23	250	21	303
Arkansas	11	254	8	303

A fair question to ask, however, is whether these regional differences are a function of traditional regional differences, or simply a result of the fact that states within a region tend to have assessment programs, and reasons for assessment programs, that are similar to each other. For example, minimum competency testing with high stakes for students has been an assessment design and rationale that has been implemented far more in the South than in the remainder of the country. Therefore, it might well be true that low standards have been set in the South because of the purpose of the testing program, and not because people in the South have lower academic standards than do people in the rest of the country. In contrast, the states for which Advanced Systems has been the contractor tend to be those who are in the initial stages of "educational reform," with programs that came out of legislation that included language such as, "Schools shall expect a high level of achievement of all students" (Kentucky Education Reform Act of 1990). These states also do not have high stakes for students, and plan to evaluate primarily on the basis of their *improvement*, rather than on their absolute position. All these factors make it reasonable and appropriate to set high standards. For example, if schools are being evaluated on the basis of their improvement, it makes little difference if 90 percent of the students are scoring below the proficient level. In contrast, if students will not be promoted from one grade to another unless they pass an examination, there will be great consequences if even 50 percent of the students score below Proficient.

Another variable in common to all the states with high standards is their dependence on using extended open-response questions to set standards. Kentucky, Maine and Arkansas set their standards using open-response questions only. New Hampshire used primarily open-response questions; the role of multiple-choice questions in the process was quite minor.

While perhaps not surprising, it also is worthwhile to note that the correlation between the standards states have set in reading and those in math is high: .94, for the 13 states that set standards in both areas. Certainly the context and methods used for the two content areas matched within each state, but usually the people involved were different. Since each group certainly was operating without knowledge of how their standards compared to national standards, the high correlation is perhaps more surprising than it appears at first glance.

Last, but certainly not least, a point worth noting is the high degree of similarity among the standards set by New Hampshire, Kentucky, Maine, and Arkansas. Despite the fact that those states are located in very different parts of the country and used different (although related) procedures over a period of four years, they wound up with standards that were within 11 points of each other in reading, and 5 points of each other in mathematics.

Since the standards in Arkansas were set at grade 11, those standards are readily understood, and their appropriateness evaluated, by the general public. Therefore, Arkansas produced a document that shows these standards that seems to be broadly accepted in these states. That document is attached to this paper so that readers can evaluate the appropriateness of those standards for themselves.

CCSSO

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

TM027775

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: States Set Common Standards, IF......

Author(s): Richard K. Hill

Corporate Source: Advanced Systems in Measurement & Evaluation, Inc.

Publication Date: June 16,1 1997

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☑

Check here
For Level 1 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) *and* paper copy.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

The sample sticker shown below will be affixed to all Level 2 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2

□

Check here
For Level 2 Release:
Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but *not* in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: Richard K. Hill

Printed Name/Position/Title: Richard Hill/President

Organization/Address:
Advanced Systems in Measurement
& Evaluation, Inc.
171 Watson Rd., Dover, NH 03820

Telephone: (603) 749-9102

FAX: (603) 749-6398

E-Mail Address: Rich@asme.com

Date: 9/17/97

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or,* if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC Clearinghouse on Assessment and Evaluation
210 O'Boyle Hall
The Catholic University of America
Washington, DC 20064

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2d Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com